DOCUMENT RESUME

ED 175 009 CS 205 045

AUTHOR Bailey, Richard W.; And Others
TITLE Measuring Student Writing Ability.
PUB DATE May 79
NOTE 19p.; Paper presented at the Annual Meeting of the Canadian Council of Teachers of English (12th, Ottawa, Canada, May 8-11, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Freshmen; *Composition (Literary); *Course Evaluation; *Educational Research; *English Instruction; Higher Education; Program Development; *Student Improvement; *Writing Skills

ABSTRACT

As part of the development of a university-wide writing program at the University of Michigan, a study was undertaken to judge the effectiveness of various composition courses. A pretest consisting of a writing sample was given to 3,913 entering students and evaluated. Students were then assigned to a regular composition class, to a tutorial composition course, to a writing workshop or to a combination of two of these. Some students were exempted from the composition requirement. A second assessment of 398 students was undertaken after the first semester of study. The results showed that the mean score for the entire group increased, the number of students in the lowest categories fell by half, and the number of students in the two highest categories increased by nearly three times. Of the 13 subgroups of students representing various options that might foster the improvement of writing, all but three improved significantly, including those taking the standard freshman course and those who postponed electing composition to the term following the study. Those students identified through the initial assessment as needing special help improved their scores by nearly 40%, more than double the improvement shown by any other group studied. Linguistic correlates of writing ability showed that the best papers differed from the worst in number of different words used, length of essay, repeat rate of function words, and proportion of words occurring once. (TJ)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Richard W. Bailey

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED175009

MEASURING STUDENT WRITING ABILITY

Richard W. Bailey
Richard T. Brengle
Edward L. Smith, Jr.

In January 1978, the faculty of the College of Literature, Science, and the Arts at the University of Michigan voted overwhelmingly for a new proposal aimed at improving the writing of undergraduates. While no single viewpoint pervaded faculty opinion, discussion and debate revealed a widely held belief that student writing has declined markedly in recent years and that many otherwise accomplished and enthusiastic students are hampered by an inability to write effectively. Opinions variously decrying this alleged decline in literacy were voiced, sometimes with great feeling, by junior and senior faculty across the college--in the humanities, the social and the natural sciences.

The program finally adopted came only after extensive consultation among many different faculty groups. It comprises three different parts, each designed to effect major changes in the teaching of writing. The first is the reading of an impromptu essay written by all

---

Research reported here was made possible through a grant from the Ford Foundation. While many of my colleagues have been particularly helpful, I am most grateful to the persons responsible for the reading and evaluation of the writing samples, particularly Richard T. Brengle, Barbey Dougherty, Francelia Clark, Michael P. Clark, Helen Isaacson, Barbra S. Morris, Kathryn Weidner, and Frances B. Zorn.

CS205045

incoming students--new freshpersons and transfer students alike--and has an important political objective in signalling the idea that writing is important in university studies. Even though various objective tests provide variously effective predictors of writing ability, we believe that exacting a simple of their writing will itself suggest to students that writing, and not its objective correlatives, is central to the entire curriculum.

The second part of our new program involves the teaching of composition in the freshman year. For most students, the standard composition course will provide instruction and practice in a variety of rhetorical and linguistic modes, geared to a variety of purposes, and reflecting the separate demands of differing audiences. Our freshman curriculum replaces literary analysis and the exploration of social issues with respectful discussion of the various modes of writing students are likely to encounter in their advanced studies. In addition, a variety of intensive courses have been introduced to develop skills and to enhance the confidence of writers identified as unprepared for the demands of college writing.

The third part introduces an upper-division writing requirement that will oblige students to elect a course, normally taught in their major field of study, that is designed to develop more adept writing within the context of a professional specialty. To this end, we are now engaged in a series of seminars for senior faculty to assist them in designing courses that will make writing a central focus in fields ranging from geology and psychology to Far Eastern languages and history of art. In short, the Michigan faculty have accepted corporate responsibility for the teaching of writing across the curriculum.

As a first step toward implementing our new program, we began collecting samples of their writing from the new students enrolled from May 1978 to the present. Elsewhere, my colleague, Michael Clark, has already explained some of the salient points of our evaluation of these samples: our attempt to ensure uniform conditions of testing and evaluation and our lengthy training sessions for the experienced composition teachers employed to evaluate the samples. Our data suggest that the training program was effective: individual bias was reduced by regular sessions in which the papers of each student group were discussed. Further uniformity was gained through the collective formulation of a list of criteria for evaluation, and even though all readings were "holistic," the relative value of mechanical errors or rhetorical organization, for instance, was often weighed to ensure that readers would evaluate papers in the same way. Common judgments were encouraged by a regular re-pairing of the two readers who provided independent judgments of the compositions; when paired readers disagreed, a third reader was invited to serve as a referee.

Readers evaluated each composition on a scale of one (a high score) to five (a low score), but the use of various intermediate grades led to an actual scale of 0.88 (for a "one plus") to 5.15 (for a "five minus"). Computation of scores from the two--or sometimes three--readers was averaged to produce a composite score, and the mean was 2.9429 for the 3,913 papers read between May and September of 1978. These numbers take on a more human scale when we report that there were 8,414 readings during the four month period. In other words, all of the papers were read by two different raters independently and about 15% (or 586) were read a third time to settle any small differences of judgment arising

from the first two readings. Nearly all of the readings--7,590--were performed by a faithful cadre of seven readers; the rest were read by five others who attended the training seminar but fewer rating sessions. One person heroically read 1,857 papers! In short, the validity of the experiment hinged on the careful training, re-training, and hard work of a small group of diligent workers.

Once the raters had settled on their final scores, a further decision was made about the placement of the writers. The possible categories are displayed in Table I. Students might write so well as to deserve exemption from the normal composition requirement: just under 4% did so. When we recall that all incoming students were evaluated, it is somewhat disquieting to note than more than 10% enrolled after having completed a composition course at another institution. As we will later demonstrate, the disparity between the two figures may not be attributable to any general incompetence of transfer students but --in many cases, at least--to the raters' reluctance to deny introductory composition to anyone in need!

As Table I shows, the raters recommended to some writers that they visit our Writing Workshop, a facility designed to assist students in coping with particular and well-defined problems. Most often, this recommendation was made in conjunction with an assignment to a composition course, but sometimes even the "exempted" writers were encouraged to make use of the Workshop to develop some aspect of their writing skills. Students with serious deficiencies--274 from this group--were advised to enroll in special tutorial sessions, small courses carrying regular composition credit but designed to allow more intensive and frequent work between students and a selected group of experienced,

professional composition teachers. As I will demonstrate in a moment, these tutorial sessions were dramatically successful in helping deficient writers.

One obvious aspect of the data reported in Table I requires further comment: the fact that there is no inevitable correlation between the highest scores and exemption and the lowest and assignment to tutorial. The explanation lies in the conversations that took place among the raters once the final scores had been assigned. A few students were judged to be superior with separate rater scores of, for instance, "one minus" (or 1.25) and "two." In discussion, the raters sometimes took the option to exempt such a student from further work despite the combined score of 1.625. The same sort of outcome might also arise in conversations about the poorer writers, resulting in a recommendation for extra help of some kind even when the strict application of the final composite score might not point unambiguously in that direction. In general, however, raters agreed, not only in assigning scores, but also in the recommendation.

Evidently enough, some special circumstances intervened between the rating session and the placement of students. A few writers, who persuaded our staff that their skills were not accurately measured by the assessment, gained exemption from the requirement. Others learned of the opportunity for special help offered by the tutorial and requested permission to enroll in one of those sessions instead of the standard composition course. Nonetheless, both sorts of requests were rare: students seldom argued that they had been unjustly treated, and the attempt to place students by their level of ability seems to have been exceptionally successful.

In order to measure the direction and degree of change in writing ability that took place during the first semester of study at the University, we administered a second assessment to more than 1,400 students during December 1978. While it might have been desirable to have rated all of the papers collected, our readers were unwilling to undertake such an onerous task, particularly when the outcome has value for "research purposes" but little immediate value for students. Therefore, a carefully selected sample of 398 students (just 10% of the total assessment group) was chosen in an effort to evaluate whatever change in writing ability had taken place during the fourteen weeks of the semester. Since we were particularly eager to trace the behavior of the group regarded as the least skillful during the summer reading, that group was somewhat over-represented in the sample, but otherwise all groups in the assessment were proportionately represented. Every effort was made during this evaluation session to ensure that the important conditions observed during the assessment reading were replicated: the initial calibration session was followed by independent judgments of variously paired readers. Papers were presented in a way that made it impossible to recognize their authors, and raters were warned that some papers in the sample were written by students destined for composition or special help but not yet enrolled in courses where it might be obtained.

Table II shows the score distribution for the sample of 398 students studied. Improvement is obvious. Though the mean score assigned to the whole group rose only half a point on the average (from 3.0824 to 2.548), the numbers in the lowest categories fell by half, while the numbers in the highest two categories increased by nearly three times.

Such a change is highly significant from a statistical point of view. There are still many students of middling ability, of course, but the direction of the trend is unmistakable.

Within the sample of 398, students were chosen to represent various options that might variously foster the improvement of writing. Of the thirteen sub-groups, all but three improved significantly (i.e., showed a change in scores significant at the .05 level or better when the T-test is applied): one group--transfer students--showed no significant improvement by this measure. But this group presented good scores on entrance (ranking third in writing skill) and their improvement was limited by the fact of their original high place on the scale. A second group failed to improve at a statistically significant level--the group originally exempted by the summer assessment. But, like the transfer group, these students were originally at the top of the scale --with a group mean of 1.3.--and thus had no room to move upward in the post-test. A third group in whose post-test writing no significant improvement was observed represents one of the alternative options to composition offered by the University; this program is now undergoing careful and critical review.

Of the ten groups that improved significantly, three merit particular attention. To everyone's satisfaction, the group representing the standard freshman course improved from a mean of 3.0 to a post-test score of 2.5. Did Freshman English make a difference? When we examine the group of students who postponed electing composition to the winter term, we find that they improved their scores at almost exactly the same rate. In both instances, about two-thirds of the students received higher scores in December and about one-quarter received lower scores

(the rest had the same score on both occasions). Several explanations for this result are possible. One is that students will improve just as much through the writing they are obliged to do in other courses as they do under the tutelage of the standard course. Distasteful as this conclusion must be to teachers of Freshman English, it will inevitably be a central concern of our future investigations. Another hypothesis is that our sample of writers representing the postponed class is somehow biased, and it is entirely possible that such is the case. Students who provided writing samples for us from this group were induced to do so by the payment of an $8.00 honorarium for their work. Even though the same number of poor writers (as measured by the assessment during the summer) were selected for study as in the Freshman English group, it may well be that the genuinely inept declined the opportunity to earn a small sum of money for a distasteful task. Their counterparts enrolled in Freshman English did not find it so easy to escape our scrutiny since the post-test was administered during a regular class hour. A study now underway to measure assessment and post-test scores against grade average, proposed field of study, high school class rank, and similar demographic variables will enable us to identify a possible bias in our sample group.

The third of the groups showing improvement that calls for special comment is the cadre of students identified through the assessment as in need of special help. Of all the groups selected for study through the post-test, this one had the lowest average score on the summer sample: 4.48. Of the 36 students chosen to represent this group, 34 achieved higher scores in December and the group mean on the post-test was 2.78, nearly the same as the final score for the students who

elected composition. As measured by our raters, these students improved their scores by nearly 40%, an improvement more than double that of any other group studied. Such an outcome is the most effective evidence for the hard work of both students and teachers in the tutorial sessions. Surely it is the poor performance of students like these that most concerns the faculty. This group of students deserves our diligent efforts, and our results clearly demonstrate that these efforts have not been wasted.

Now that we have completed the holistic evaluation of assessments and post-tests, we are embarked on an analysis that will correlate writing ability as measured by these instruments with a whole variety of student "characteristics." As might naturally be expected, we have already found that students who do well on the advanced placement composition test and who achieve high scores on the verbal portion of the Scholastic Achievement Test also do well on the University of Michigan assessment, but we are also eager to test the conventional beliefs about the relation of writing skill to course grades, sex, age, field of study, and other similar variables. With the cooperation of several large school districts in Michigan, we hope to identify the types of secondary programs that seem to produce the best writers and to use the results in an attempt to improve high school writing curricula across the region. Since several of these districts maintain files of student writing, we also expect to be able to trace the development and growth of writing ability over a longer period than has yet been attempted by other investigations of the acquisition of writing skills.

In order to develop effective writing programs, we are investigating the linguistic variables that correlate with writing skills as

measured by our instrument. Our eventual concern is to specify the inventory and deployment of cohesive devices in the writing samples, but we are also eager to compare essays in terms of quantitative variables and measures of syntactic variety and complexity. The first results of this effort are displayed in Table III, a study of over 16,000 words of writing from our assessment--sixteen papers by our best writers (7,922 words long in the aggregate) and twenty-three papers by our poorest writers (8,044 words long).

The results in Table III do not confirm some of the conventional wisdom about good and poor writing. No significant difference emerges from a comparison of the length and variety of sentences in the two groups. Both groups are nearly identical for these measures, and it is our impression that the poorest writers use their short and long sentences in ways that reveal considerable syntactic sophistication but have an impoverished awareness of the ways in which sentences fit together to form larger patterns of discourse. Vocabulary measures show that the better writers have a much larger inventory of words in their essays--an artifact, certainly, of the fact that the best papers are considerably longer than the poorest ones. Surprisingly, hardly any of the student writers made use of all the time available for the exercise; the length difference, then, may be related to the students' conception of the requirements of the task and their comfort with it. Each group took as long as "necessary" to complete the essays, but the poorest writers apparently were satisfied with having less to say about their subjects. The difference here is of "amplitude." Writers judged superior provided many examples and illustrations and explored a greater range of possible viewpoints. Poorer writers were satisfied with less,

and exhausted their subjects (if not themselves) after a relatively superficial treatment of the topics.

In developing their essays, both sets of students repeated content words at about the same rate and thus the words that carry the burden of meaning--nouns, verbs, adjectives, and adverbs--appear to be linked through repetition in about the same way in the two groups. But the superior writers had a greater variety of vocabulary as measured both in absolute terms and by the percentage of words occurring once, the hapax legomena. Since the hapax are more frequent in the best writers, we can see that the results indicate a theme--measured by the repetition of core content vocabulary--balanced against the variations of synonyms and expanded paraphrase. For the repeat rate of content words to be stable for each group, the superior writers had to use their thematic content vocabulary more frequently than the poorer ones at the same time that they were using their "variation" vocabulary to obtain greater range of expression. The same sort of conclusion is supported by a comparison of the type/token ratios for the two groups, measured here on a logarithmic scale to eliminate the effect of differing sample sizes. Both groups repeated vocabulary at about the same rate, but the superior writers began with a larger vocabulary of both function and content words. Superior writers, then, are characterized by using words at the extremes of the frequency scale--more words used just once and more words used again and again. Poorer writers, using fewer different words, missed the opportunity to link ideas and structures through well-managed repetition and failed to introduce variety by using alternative expressions.

Even though only four of the eight variables in Table III show a statistically significant difference between the best and poorest writers--as measured by the T-test--all of the variables taken together show that the groups differ very markedly indeed--at .0006 with an F-statistic of 4.93. Few would be surprised by this result. But the nature of the differences is, I believe, suggestive. The power of the superior writers comes not from the capacity to write long and complex sentences or in their ability to control the mixture of short and long sentences: both groups are nearly the same in this respect. The difference lies in the differing skills that are measured by amplitude and variety across sentence boundaries and that show the ability to take a subject in its "wholeness" and explore it.

While our research is still very much in progress, I believe that we can draw several preliminary conclusions of general interest. Our approach to testing appears to be effective: students have been able to display their skills and our raters are able to form uniform and reliable impressions. The growth and change in writing ability appears to be directly correlated with the intensity and frequency of writing tasks during the first semester of university study. The writings themselves suggest that students come to Michigan with a good mastery of skills at the sentence level and below, but that they differ significantly in their ability to manage discourse about a subject by exploring its variety in a sustained and "ample" piece of exposition. If these conclusions are confirmed in further research, teachers will want to make curricular changes that will increase our ability to help students learning to write.

Richard W. Bailey, "Correlates of Student Writing Ability"

Table I: Distribution of Assessment Scores, May-September 1978

Scores: (High)	.88-1.50	1.51-2.25	2.26-3.0	3.01-3.75	3.76-4.25	4.26-5.15 (Low)
Placement						
Exempted (3.45%)	2.81	0.64	--	--	--	--
Exempted with Writing Workshop (0.33%)	0.18	0.15	--	--	--	--
Composition (76.24%)	1.07	13.00	43.61	15.99	2.43	0.13
Composition and Writing Workshop (13.11%)	--	0.20	3.53	5.88	2.78	0.72
Tutorial (6.82%)	0.05	--	0.03	0.08	0.74	5.93
Missing data (0.05%)	--	--	0.05	--	--	--
Total (N=3,913)	4.11	13.99	47.22	21.95	5.95	6.78

Table II: Distribution of Scores of Post-Test Population

Scores: (High)	.88-1.50	1.51-2.25	2.26-3.0	3.01-3.75	3.76-4.25	4.26-5.15 (Low)
Assessment scores (all students)	4.11	13.99	47.22	21.95	5.95	6.78
Assessment scores (post-tested group. N=398)	4.27	10.30	45.23	21.11	6.28	12.81
Post-test scores (post-tested group)	18.09	20.60	36.93	14.07	6.28	4.02

Table III: Linguistic Correlates of Writing Ability

	"Ones" (Best Papers)	"Fives" (Poorest Papers)	Significance
Average Sentence Length	20.71	19.96	--
Variety of Sentence Lengths (std. deviation of S. L.)	5.63	5.75	--
Number of different words	245.88	170.17	.0001
Length of essay	499.50	349.74	.0014
Repeat rate of content words	1.4	1.49	--
Repeat rate of function words	3.65	3.04	.0085
Proportion of hapax legomena	.696	.658	.0406
Log Type/Token ratio	.886	.882	--